UCRL-CONF-211202

LAWRENCE LIVERMORE NATIONAL LABORATORY

Bioinformatics for Diagnostics, Forensics, and Virulence Characterization and Detection

S. Gardner, T. Slezak

April 8, 2005

DHS Homeland Security Conference
Boston, MA, United States
April 27, 2005 through April 28, 2005

Authors: Shea Gardner, Tom Slezak
April 4, 2005
Manuscript for the DHS Homeland Security Conference April 27-28, 2005

Bioinformatics for Diagnostics, Forensics, and Virulence Characterization and Detection

We summarize four of our group's high-risk/high-payoff research projects funded by the Intelligence Technology Innovation Center (ITIC) in conjunction with our DHS-funded pathogen informatics activities. These are ***1) quantitative assessment of genomic sequencing needs*** to predict high quality DNA and protein signatures for detection, and comparison of draft versus finished sequences for diagnostic signature prediction; ***2)*** development of ***forensic software*** to identify SNP and PCR-RFLP variations from a large number of viral pathogen sequences and optimization of the selection of markers for maximum discrimination of those sequences; ***3)*** prediction of signatures for the ***detection of virulence, antibiotic resistance, and toxin genes and genetic engineering markers*** in bacteria; ***4)*** bioinformatic ***characterization of virulence factors to rapidly screen genomic data*** for potential genes with similar functions and to elucidate potential health threats in novel organisms. The results of (1) are being used by policy makers to set national sequencing priorities. Analyses from (2) are being used in collaborations with the CDC to genotype and characterize many variola strains, and reports from these collaborations have been made to the President. We also determined SNPs for serotype and strain discrimination of 126 foot and mouth disease virus (FMDV) genomes. For (3), currently >1000 probes have been predicted for the specific detection of >4000 virulence, antibiotic resistance, and genetic engineering vector sequences, and we expect to complete the bioinformatic design of a comprehensive "virulence detection chip" by August 2005. Results of (4) will be a system to rapidly predict potential virulence pathways and phenotypes in organisms based on their genomic sequences.

~50 word abstract:

Designing signatures for pathogen diagnostics and forensics demands that there be sufficient genomic data and a computational infrastructure to rapidly process them. This talk will outline the bioinformatics of 1) a system to assess such sequencing needs, 2) a new forensic pipeline and its predictions for variola and FMDV, and 3) signature prediction for detecting virulence, antibiotic resistance, and genetic engineering.

Quantitative assessment of genomic sequencing needs

Our group uses computational methods to develop DNA diagnostic signatures, which are short sequences that are sufficient to uniquely identify a pathogen species [1-3]. After laboratory screening and validation, many of the signatures we generated are in widespread use by BioWatch and various federal and state agencies for detecting pathogens [4, 5]. In addition, our group has developed a protein signature pipeline that predicts peptide targets which may be developed in the laboratory as targets of antibody, ligand, or peptide binding for detection assays and therapeutics, or as targets for vaccine development [6, 7]. A question that arose from this work was "How many complete genome sequences do we need in order to predict high quality DNA and protein signatures?" To address this issue, we built a system called the Sequencing Analysis Pipeline (SAP) to guide decisions regarding the amount of genomic sequencing required to develop diagnostic DNA and protein signatures [6, 8]. We used our existing DNA and protein diagnostic signature prediction pipelines, which select regions of a target species genome or proteome that are conserved among strains of the target (for reliability, to prevent false negatives) and unique relative to other species (for specificity, to avoid false positives). We performed simulations, based on existing sequence data, to assess the number of genome sequences of a target species and of close phylogenetic relatives ("near neighbors") that are required to predict diagnostic signature regions that are conserved among strains of the target species and unique relative to other bacterial and viral species. We focused our analyses on viruses because there were a sufficient number of complete genomes available for many species in order for our simulations to generate informative predictions.

We were able to make some generalizations for DNA signature prediction [8]: For DNA viruses such as variola (smallpox), three target genomes provide sufficient guidance for selecting species-wide DNA-based signatures. Three near neighbor genomes are critical for species specificity. Most RNA viruses, which show much more sequence variation across isolates/strains than do DNA viruses, require at least four target genomes and no near neighbor genomes, since lack of conservation among strains is more limiting than uniqueness. In all cases, the best signature results occur when the genomes are chosen widely across the dimensions of time, space, virulence, etc. to maximize genetic variability. Emerging viruses such as SARS and Ebola Zaire are exceptional, as additional target genomes currently do not improve predictions, but near neighbor sequences are urgently needed. Emerging viruses often show little sequence variation, which makes sense in light of the short evolutionary time over which the species has had opportunity to evolve. Our results also indicate that double stranded DNA viruses are more conserved among strains than are RNA viruses, since in most cases there is at least one conserved DNA signature candidate for the DNA viruses and zero conserved signature candidates for the RNA viruses.

Variola virus

~98% of the variola genome is conserved, but only 4% is conserved+unique, and thus suitable for diagnostic signatures. Without other Orthopox sequences, it appears that 60% of the variola genome is unique, a tremendous overestimation.

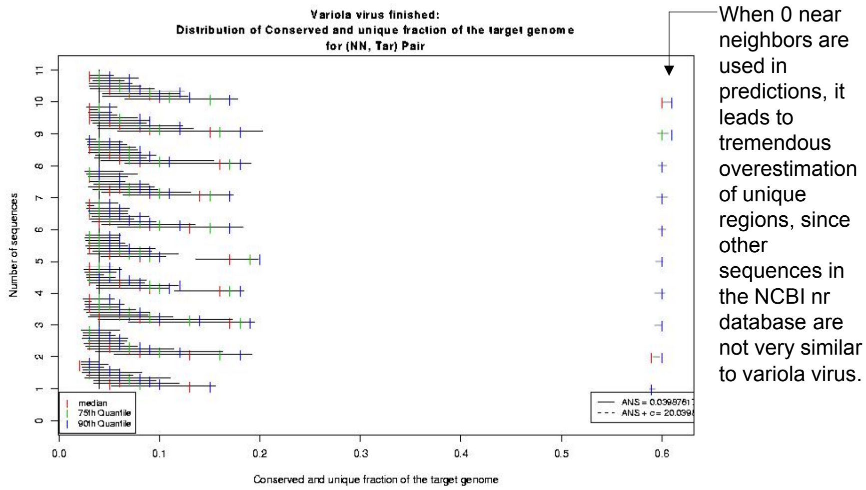

Figure 1: Range plot as described in [8], showing that the sequence data from close phylogenetic relatives is essential in eliminating non-specific regions of the variola genome to be considered for diagnostic signatures. Levels of intra-specific variation is relatively low, typical of double-stranded DNA viruses compared to RNA viruses, so that only 2 or 3 variola genomes would have been adequate to identify conserved regions of the genome suitable for DNA diagnostic signature prediction.

The number of complete genome sequences required for protein signature prediction, in contrast to that for DNA signature prediction, is highly dependent on the species under consideration, and no generalizations can be made about genome structure, such as whether the virus is RNA or DNA, or single or double stranded. We conclude that it is necessary to use the SAP as a dynamic system to assess the need for continued sequencing for each species individually, and to update predictions with each additional genome that is sequenced. One generalization that does arise, however, is that there are more protein than DNA signatures, a consequence of greater conservation at the protein than at the nucleotide level. This protein conservation is particularly notable for some divergent RNA viruses of biothreat concern. For Marburg virus, Venezuelan equine encephalitis virus, and foot and mouth disease virus there is not a single non-degenerate TaqMan DNA signature that is conserved among all strains, but there are multiple protein signatures. The fact that we can identify highly conserved, species-specific peptides indicates that these peptides, or the proteins on which they reside, may be important targets for therapeutics and vaccines, and we welcome empirical collaborations to examine these peptides in the laboratory.

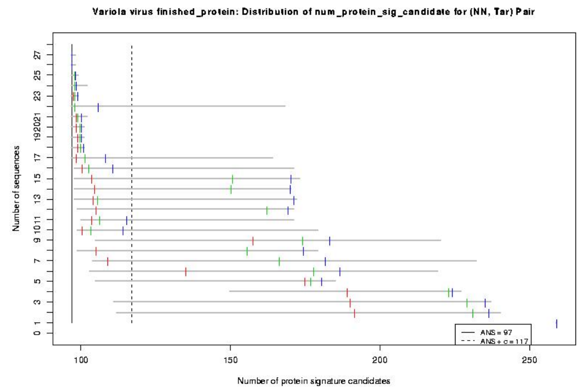

Figure 2: Range plot as described in [6] for protein signatures prediction. Data indicate that in the range from one to ten sequences, additional sequencing was useful to eliminate non-conserved peptide signature candidates, but thereafter, particularly beyond 17 genomes, additional sequences do not continue to eliminate peptide signatures, as all variable regions of the proteome have likely been identified.

We also use SAP to assess whether whole genome draft data is sufficient or whether finished sequencing is required to predict DNA and protein signature candidates [9]. We used actual draft *Marburg* virus sequence data, and we simulated *variola* virus draft from finished sequences, imposing a range of error rates to mimic several levels of sequencing quality or coverage. Simulations indicate that intermediate to high quality draft with error rates of 10^{-3}-10^{-5} (~8x coverage) of target organisms is suitable for DNA signature prediction. Low quality draft with error rates of ~1% (3x to 6x coverage) of target isolates is inadequate for DNA signature prediction, although low quality draft of close phylogenetic relatives, for eliminating non-specific candidates, is sufficient, as long as the target genomes are of high quality. For protein signature prediction, sequencing errors in target genomes substantially reduce the detection of amino acid sequence conservation, even if the draft is of high quality.

This past year, the National Interagency Genomics Sciences Coordinating Committee (NIGSCC) determined sequencing priorities and planned investments, assisted by a report we prepared detailing SAP bioinformatic predictions. The report assessed needs for sequencing additional strains of target pathogens and their close phylogenetic relatives for all sequenced pathogens on the CDC's Category A-C biothreat list.

Forensic software

Microbial forensics and epidemiology is important in tracking the source of a pathogen, whether the disease is a naturally occurring outbreak or part of a criminal investigation. We developed software called SPR Opt (**S**NP and **P**CR-**R**FLP **Opt**imization) to automate the process for multiple, complete genomes of identifying all SNP and fragment length polymorphisms, clustering those variations into haplotypes, and determining the maximal level of discrimination that can be obtained among those genomes using the fewest tests [10]. The PCR-RFLP analysis includes prediction and selection of optimal primers and restriction enzymes to enable maximum isolate discrimination based on sequence information.

These analyses highlight variable regions based on existing sequence data. However, these markers may be heterogeneous among unsequenced isolates as well, and thus may be useful for characterizing the relationships among unsequenced as well as sequenced isolates.

This is the first software to optimize the selection of forensic markers to maximize the information gained from the fewest assays, accepting whole or partial genome sequence data as input. As more sequence data becomes available for multiple strains within species, automated, computational tools will be essential to make sense of large amounts of information and to guide efforts in the laboratory. SPR Opt is freely available for non-profit use, and can be downloaded at http://www.llnl.gov/IPandC/technology/software/softwaretitles/spropt.php.

The SNPs identified by the current version of the software are defined in the strict sense of single nucleotide polymorphism: that is, a SNP is a single base position that varies across the input genomes and is surrounded by sequence that is conserved across all those input genomes. The length requirement of the conserved sequence upstream and downstream of the variable base in order for the base to be considered a SNP is set by the user. The conserved sequence must occur once, and only once, in all the genomes under consideration. In most of our test cases, we have found that a length of 5-7 conserved bases on either side of the SNP maximizes the number of SNPs that can be found and the level of discrimination possible among the genomes; shorter than this and sequences are often repeated within a given genome, longer than this and sequences are often no longer conserved across all the input genomes. A length of 12-mers on either side of a variable position allows one to identify oligo probes suitable for laboratory testing on platforms such as microarray chips. A mark of the discriminatory power of SNPs is the number of unresolved clusters into which the genomes can be grouped. If the number of unresolved clusters equals the number of input genomes, then isolate-level discrimination is possible.

We used this software to analyze a large number of variola genomes that were publicly available or provided to us by our collaborators at the CDC. We also performed analyses on >60 Orthopox genomes available to us at the time of our analyses, to identify SNPs that were unique to a given species within the genus. For Orthopox, we found 2087 SNPs or 920 SNPs (conserved 7-mers or 12-mers, respectively, surrounding the variable SNP base), capable of discriminating the sequences into 50 or 48 unresolved clusters (conserved 7-mers or 12-mers, respectively). Fewer than 60 probes would be necessary to determine how an uncharacterized Orthopox sample compared with the already-sequenced isolates. There are species-specific SNPs for each of the Orthopox species in the analysis (Table 1).

Table 1: Species specific SNPs within Orthopox

Species	Number Genomes	Number species-specific SNPs with conserved 7-mers surrounding SNP
buffalopox	1	4
taterapox	1	60
rabbitpox	2	9

camelpox	2	110
ectromelia	2	257
monkeypox	7	164
cowpox	3	4
vaccinia	6	1
variola	>40	175

Focusing on the variola data alone yields 1175 SNPs (with conserved 12-mers surrounding the SNP position), clustering into 220 haplotypes, and resolving the genomes into >40 clusters. It would require less than 50 probes, each with a length of 25 bases and the central base a SNP, to determine how an uncharacterized variola sample clusters with the already-sequenced isolates. Our colleagues at the CDC are using these variola and Orthopox SNPs to build a microarray to test our SNP predictions and experimentally characterize unsequenced samples. More details will be forthcoming in a future publication (Li, Gardner, et al., in prep.). Results of our collaboration examining sequence variation among variola strains was included in a report by our ITIC sponsor prepared for the White House.

For the SNP analyses of the 126 unique FMDV genomes available in GenBank in the summer of 2004, there was insufficient conservation across serotypes to find regions conforming to our strict definition of a SNP as having a variable base surrounded by conserved bases up- and down-stream. In fact, there was only one fragment longer than 10 bases conserved across all FMDV genomes (12 bases in length). Therefore, we subdivided the genomes into the 7 serotypes, and performed the SNP analyses separately on each serotype. The implication of this is the assumption that the serotype of a sample should be determined prior to running the SNP analyses that are specific to that serotype, for example, by including serotype-specific probes on a microarray, as well as the probes to discriminate the isolate-level SNPs within each serotype. Requiring conserved 5-mers up- and down-stream of the SNP base, it is possible to obtain genome-level discrimination for serotypes SAT 1, SAT 2, and SAT 3, and nearly this level of discrimination for serotypes C and Asia (Table 2).

Table 2: Summary of FMDV SNP Analyses

Serotype	Number Genomes	Number of SNP positions	Number of Haplotypes	Number of Unresolved Clusters
A	47	25	54	31
O	42	36	72	27
C	8	177	57	7
SAT 1	9	26	28	9
SAT 2	5	39	18	5
SAT 3	4	183	12	4
Asia	7	141	44	6

For serotypes A and O, for which there are more than 40 genomes each, a number of the genomes cannot be discriminated with the strict definition of SNPs. To obtain higher power of discrimination, it is necessary to loosen the strict definition of a SNP, no longer requiring that the sequence surrounding the SNP position be conserved among all the input genomes: multiple polymorphic nucleotides or indels in close proximity to the SNP position are allowed, and only a subset of the input genomes needs to align in the region surrounding the SNP.

While this loose definition of a SNP uncovers many more SNP positions, and thus a higher level of discrimination, than the strict requirements for conservation surrounding the SNP among all the genomes, it is more difficult to take such loosely defined SNPs to the lab for testing. For example, for a given SNP, oligo variants must be included corresponding to the surrounding variations as well as the target SNP position. For each of FMDV serotypes A and O, variations characterized by loosely defined SNPs enable all but 2 of the genomes to be uniquely discriminated (all but a22iraq64 iso86 and a22iraq70 iso92 for serotype A, and SKR/2000 and o1skr iso85 for serotype O). While there were no SNPs according to the strict SNP definition for all FMDV serotypes analyzed together, nearly 3000 SNPs were uncovered based on the loose definition of a SNP, enabling the 126 genomes to be separated into 118 unresolved clusters, or near-isolate-level discrimination. From this same analyses, for each of the FMDV serotypes except SAT 2 and A, there were one or two SNPs that were serotype specific (the same nucleotide identity in all genomes of that serotype, and not in genomes of other serotypes). Surprisingly, most of these serotype-specific SNP positions were not located in the P1 antigenic region that determines vaccine selection.

A potential application of SNP analysis is to reconstruct evolutionary relationships of uncharacterized strains relative to the isolates that have already been sequenced by empirically determining the base identities at each position known to be a SNP among the sequenced strains. Although such phylogenetic characterization based on SNPs may correspond generally with a more accurate phylogeny based on full genomes, there are potential errors introduced in such analyses (Figure 3A,B). For example, such an analysis of FMDV serotype O SNPs suggests that the o11indonesia iso52 strain appears most similar to the orey-iran iso53, although the maximum likelihood phylogeny based on a full genome alignment indicates that the Indonesia strain is more closely related to several South American strains.

We welcome collaborations for testing these computational FMDV SNP data in the laboratory.

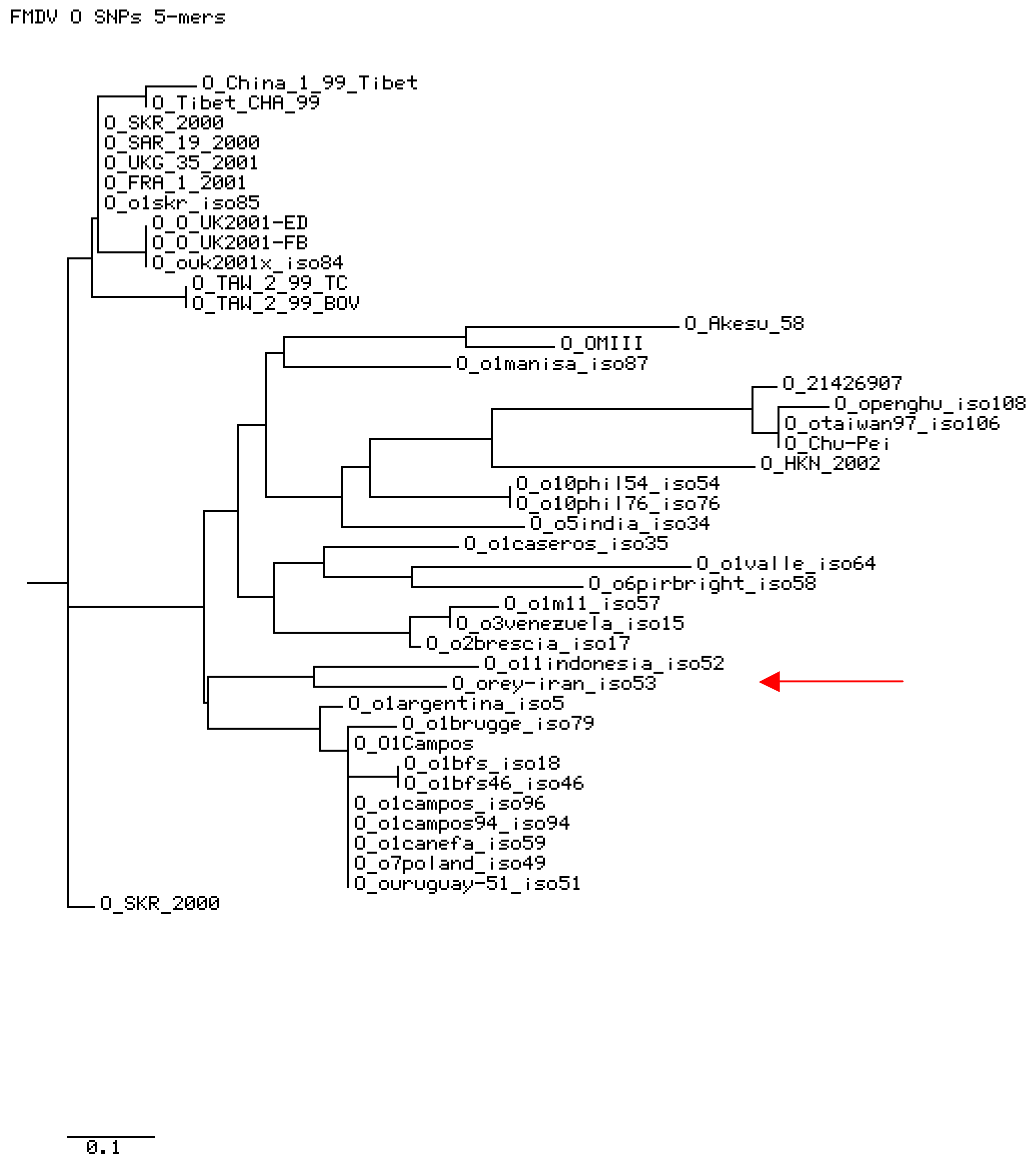

Figure 3A. Tree based on SNPs (a SNP matrix), using the dnadist (maximum likelihood) and neighbor methods in PHYLIP (http://evolution.genetics.washington.edu/phylip.html) and visualized using the Phylogenetic tree printer (http://iubio.bio.indiana.edu/treeapp/treeprint-form.html).

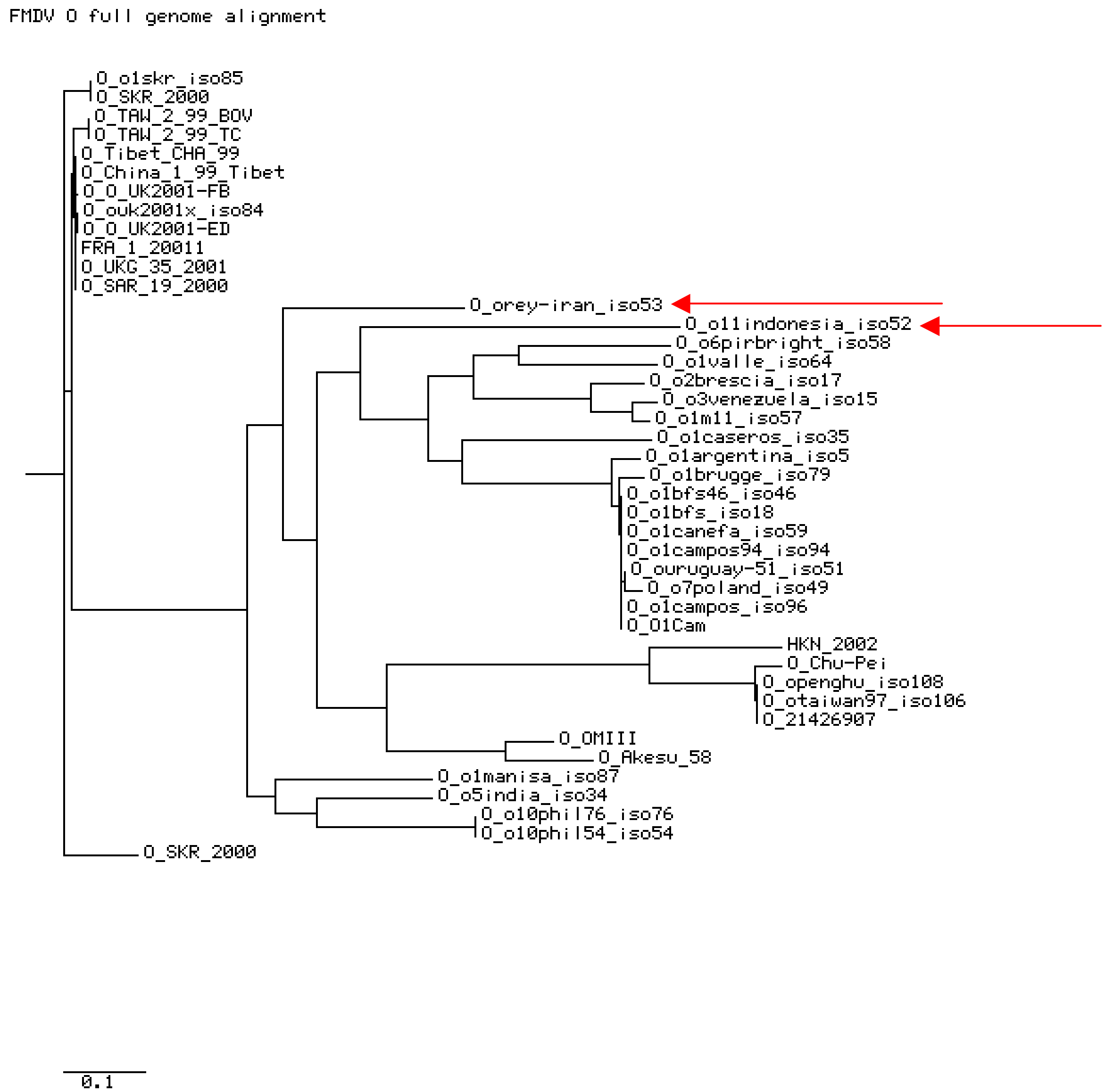

Figure 3B. Tree based on full genome alignment.

Virulence Detection

We are developing functional signatures to detect virulence and antibiotic resistance, selecting probe sequences that could be used on a virulence detection microarray chip. The goal is to detect what a pathogen can do, not only what species it is. Toxins, antibiotic resistance, genes involved in known virulence pathways, and vector sequences for genetic engineering are included. If built as we would plan, the chip would also contain a number of species-specific probes. Such a chip would rapidly indicate, in a highly multiplexed fashion, potential virulence factors in a sample, the species present, and whether there is evidence of bacterial genetic engineering.

A number of taxonomically divergent species might share a similar virulence mechanism due to the horizontal transfer of genetic information, for example, on plasmids or pathogenicity islands. A virulence gene (or entire gene "kit") may evolve independently in several species, but still exhibit conserved protein sequence or virulence function, resulting in potentially high levels of DNA sequence variation. An excellent example of this is the Type III secretion mechanism, whose evolution can be viewed graphically at http://www.genome.jp/kegg/ortholog/tab03070.html). Thus, given a single gene template, there may be many homologous gene sequences that a pathogen chip should also detect. Automated software has been developed to select all sequences in NCBI GenBank that are homologous to the template based on BLAST results of all hits with a minimum percent identity over a minimum fraction match length of the template (currently 50% and 0.7, respectively). All sequences are gathered that match according the specified criteria, including only the matching portions of genomes or other larger NCBI entries that have a region homologous to the template.

Next, sets of probes are developed that in combination should detect all of the homologous genes for a given template. If there is a conserved region as long as the desired probe among all the homologous sequences, that meets all probe requirements (minimal secondary structure, appropriate T_m, no self dimerization), then a single probe could be capable of detecting all the homologues. However, in most cases examined, several probes are required to detect all sequence variations. Software has been developed to find all possible probes in the input sequences that meet the length, T_m, and lack of secondary structure and dimerization requirements, and then from these to select a minimal set of probes that should, in combination, detect all the homologous sequences. Once such a set has been selected, the probes are BLASTed against the NCBI nucleotide non-redundant (nr) database, so that those with potential undesired cross reactions (e.g. matches to human or other DNA that is unlikely to be a virulence sequence) can be avoided.

Using this software, 25-mer probe sets have been predicted for more than 200 virulence genes, requiring approximately 1500 probes with specificity to approximately 4500 genes involved in virulence pathways, antibiotic resistance, toxin production, or that suggest the presence of vectors for genetic engineering. We continue to add to this collection, and anticipate having the preliminary bioinformatic design of a virulence detection chip ready by the end of summer, 2005. We hope to work with our sponsors to take this chip to the lab testing stage in FY06.

Virulence Gene Discovery

Kathryn Swan, a postdoctoral fellow in our group, has begun to use computational tools to identify and characterize genes and functional gene sets capable of conferring virulence on an organism. The purpose of this work is to identify potential virulence factors, virulence pathways, toxins, and genes associated with antibiotic resistance through patterns of sequence similarity to know genes, and then to use the patterns that are discovered to computationally predict potential virulence pathways in sequenced organisms that have not yet been characterized. These tools may quickly elucidate potential health threats in novel organisms. The technical approach is to gather existing and create new Hidden Markov Models for virulence gene families, and to describe virulence pathways by sets of HMMs. A tailored collection of these models will be created for rapidly screening genomic data for potential genes with similar functions. Early results from this work appear promising, and will be tested on some pathogen isolates that have recently been sequenced and made available to us.

Conclusion

Our efforts in pathogen bioinformatics under the sponsorship of ITIC, leveraging our years of infrastructure developed under DOE/CBNP and now DHS funding, have been outlined. In turn, we anticipate that future DHS projects will draw from the tools developed and data gathered and analyzed with ITIC support. We welcome empirical collaborations to validate and apply our bioinformatic results, and to help us to guide future bioinformatic efforts to best serve the goals of homeland security in terms of validated diagnostic and forensic assays.

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48. This work was supported by the Intelligence Technology Innovation Center.

1. Fitch JP, Chromy BA, Forde CE, Garcia E, Gardner SN, Gu P, Kuczmarksi TA, Melius C, McCutchen-Maloney SL, Milanovich FM *et al*: **Biosignatures of pathogen and host**. In: *Proceedings of the IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS): 2002; Raleigh, NC*; 2002.
2. Fitch JP, Gardner SN, Kuczmarski TA, Kurtz S, Myers R, Ott LL, Slezak TR, Vitalis EA, Zemla AT, McCready PM: **Rapid Development of Nucleic Acid Diagnostics**. *Proceedings of the IEEE* 2002, **90**(11):1708-1721.
3. Slezak T, Kuczmarski T, Ott L, Torres C, Medeiros D, Smith J, Truitt B, Mulakken N, Lam M, Vitalis E *et al*: **Comparative genomics tools applied to bioterrorism defence**. *Brief Bioinform* 2003, **4**(2):133-149.
4. Heller A: **BASIS counters airborne bioterrorism**. *Science and Technology Review* 2003:http://www.llnl.gov/str/October03/pdfs/10_03.02.pdf.
5. Morris T: **LRN results messenger BioWatch deployment**. *http://wwwcdcgov/phin/conference_presentations/05-14-03/5E/2003%20PHIN%20Conference%20Session%205E%20-%20Tim%20Morrispdf* 2003.
6. Gardner SN, Kuczmarksi TA, Zhou CE, Lam MW, Slezak TR: **A System to Assess Genome Sequencing Needs for Viral Protein Diagnostics and Therapeutics**. *Journal of Clinical Microbiology* 2005, **In press**.
7. Zhou CE, Zemla AT, Roe D, Young M, Lam M, Schoeniger JS, Balhorn R: **Computational approaches for identification of conserved/unique binding pockets in the A chain of ricin**. *Bioinformatics* 2005, accepted pending revision.
8. Gardner SN, Lam MW, Mulakken NJ, Torres CL, Smith JR, Slezak TR: **Sequencing needs for viral diagnostics**. *Journal of Clinical Microbiology* 2004, **42**(12):5472-5476.

9. Gardner SN, Kuczmarski TA, Lam MW, Mulakken NJ, Smith JR, Torres CL, Zhou CE, Slezak TR: **Draft versus finished sequence data for diagnostic signature development**. *Submitted* 2005.
10. Gardner SN, Wagner MC: **Software for optimization of SNP and PCR-RFLP genotyping**. *BMC Genomics* 2005, **submitted**.